PERCEPTUAL SPACES AND THE IDENTIFICATION OF NATURAL AND SYNTHETIC SENTENCES

N. BACRI *
Laboratoire de Psychologie Expérimentale

C.N.R.S. - E.H.E.S.S.
54, bd Raspail 75006 PARIS

Is synthetic speech just degraded speech or is it processed as a specific perceptual space? The identification responses to 8 phonetically balanced lists of ten sentences each, using several syntactic structures, were studied for four sets of stimuli (natural speech, LPC speech, synthesis by diphones using two text-to-speech systems). All the stimuli were intensity equalized, then degraded by a masking pink noise. Phonetic and prosodic cues effects were strong, while the effect of syntax was weak. The choice of sentence identification strategy depends on the natural vs synthetic nature of the speech used and on SNR: a step-by-step decoding for impoverished synthetic speech and a SNR below 8 dB, backward lexical interpretation for natural speech or a low noise. Acoustic cues redundancy and masking noise level impose the choice of specific cognitive processing modalities.

In the case of spoken language, sentence perception and comprehension imply the interaction of both acoustic and linguistic sources of knowledge to identify word boundaries, select word candidates and construct a meaningful sequence. According to identification tasks using a gating paradigm in which signal duration is varied /3, 4, 9/, data support the assumption of a parallel and interactive processing of acoustic-phonetic information and of syntactic-semantic information provided by the sentence context. It is the redundancy of lower-order and higher-order sources of information which can explain the listener's ability to understand speech even under degraded conditions. But the redundancy of acoustic-phonetic cues by themselves is also of importance. It is possible to evaluate its weight by comparing sentence recognition performance for natural speech and synthetic speech of different qualities.

Previous research has demonstrated that synthetic speech is more difficult to recognize than natural speech /8/. This is perhaps due to what Nusbaum and Pisoni /7/ call the "noisy speech" hypothesis i.e. the fact that acoustic structure of synthetic speech is somehow degraded, as is the case of natural speech in noise. But according to the "impoverished speech" hypothesis, the rather bad performance for synthetic speech corresponds to a specific cognitive processing. Listeners must adapt their perceptual and identification strategies to a signal which is in its nature different from natural speech: they have to build a new perceptual space.

The present experiment aims at studying how naive listeners, without a previous knowledge of synthetic speech, can manage to understand sentences with different degrees of syntactic complexity, either natural or digitized, or generated by good vs. low-cost text-to-speech systems. Moreover stimuli were degraded by adding varying amounts of pink noise. The main hypothesis is that the level of performance and kind of errors will be linked to the quality of the sets of stimuli i.e. to the characteristics of the potential perceptual space. In any cases they will be significantly different for natural and synthetic speech. Another assumption bears on the effect of syntactic and semantic complexity. As speech becomes less intelligible, according either to its quality or to speech-to-noise ratio (SNR), listeners will rely more heavily on linguistic structure, so that easy-to-parse sentences would be better understood than less predictable ones, specifically as the quality of synthetic speech becomes worse /7/. Finally, following the researches on synthetic speech training /2/, it can be hypothesized that the results will improve from the first to the second session.

Speech materials and systems

Eight phonetically balanced lists of ten sentences each, covering a range of syntactic structures and semantic degrees of plausibility /1/, were read by a trained female speaker, with a neutral intonation and a 4.27 syllables/second speech rate. The first set

* This paper was supported by the CNET. Experiment was carried out in col. with P. Graillot, when the author was on leave at CNET, Lannion-A. She would like to thank him, C. Sorin and other members of RCP Department for their help.

of stimuli, A_1, consisted of these naturally spoken sentences. Audio tapes of the original sentences were then sampled at 16 kHz (16 coefficients), digitized by a linear prediction coder, and stored on disk by a PDP-11/34 computer. This second set of materials will be referred to as A_2. The two other sets were generated using synthesis by diphones according to two text-to-speech systems. The high-quality one, A_3, was processed with all frames set to 13 ms using a PDP-11/34 computer, and generated from a diphone dictionary recorded by a male speaker at a 3.42 syl./sec. speech rate. Prosody was a good approximation of natural speech. The last set of stimuli, A_4, was processed by a low-cost system using a diphone dictionary recorded by a female speaker at a 3.18 syl./sec. speech rate. This dictionary was implemented on a micro-processor (26 ms period). Some rough prosodic markers were added.

Procedure and subjects
Mean intensity of all the stimuli was equalized at 71/72 dB lin. The stimuli were masked by pink noise the intensity of which decreased from trial to trial. In the first trial, SNR were of - 2 dB for natural speech, + 4 dB for LPC speech and high-quality system, + 8 dB for low-cost system. These values were chosen so that no correct response could be given at the first presentation. At each of the 6 successive presentations, the level of noise was diminished by 2 dB steps for natural speech, 2 dB then 3 dB steps for synthetic speech. Four groups of 5 subjects each participated in the experiment during 2 sessions, at an interval of 5 days. All groups were given the same recognition task. Subjects had to say what they had understood after each presentation of each sentence. Order of presentation was counterbalanced, and the systems were crossed with the lists according to a latin square design. For each group the factorial design was as follows:

$$S_5 * L_8 <A_4 * D_2> * Se_{10}$$

(S: subjects, L: lists, A: systems, D: test session, Se: sentences).
Speech-to-noise ratio at the identification threshold for all the responses, correct response percentages for each list, sentence or system, perceptual confusions and SNR at the identification threshold (IT) for correct responses were analysed.

Results and discussion
An ANOVA was performed on the SNR at the IT, after the IT reached by a subject in erroneous responses in the last trial was increased by 4 dB. Overall analysis showed that all the factors had a significant effect, especially the factor Systems (F(3, 48) = 682, p<.0001). Interactions were also significant. Three major findings were obtained for post hoc comparisons. First, the main discrepancy is between natural speech and low-cost text-to-speech system, as was expected, while the weakest is between LPC and high quality text-to-speech systems (F(1, 16) = 29.88, p<.01). This last result confirms the good quality of this synthesis, as well as the basic difference between natural and coded or synthetic speech (Table I).

Systems	Mean SNR	sd
A_1	2.59	.805
A_2	10.45	1.269
A_3	12.01	1.801
A_4	15.80	2.735

Table I - Mean SNR at the identification threshold (dB) as a function of the systems. A_1: natural speech; A_2: LPC speech; A_3: high-quality text-to-speech system; A_4: low-cost text-to-speech system.

Second, the effect of syntactic-semantic differences between sentences is significant (F(9, 144) = 9.59, p<.001), but it is higher for natural speech than for synthetic speech, and is not related to the intrinsic quality of synthetic speech. For synthetic speech systems, the presence of easy-to-parse sentences does not facilitate identification, compared with less expected structures. Third, differences between sessions are significant (F(1, 16) = 13.41, p<.01), but this effect is only due to the contrast between the reality of a kind of training for the "poor" system A_4 and the lack of learning in all the other conditions (Table II).

	Sessions		
Systems	1	2	Change
A_1	2.48	2.70	+ 0.22
A_2	10.72	10.18	- 0.54
A_3	12.22	11.80	- 0.42
A_4	16.99	14.62	- 2.37

Table II - Mean SNR at the identification threshold (dB) as a function of systems and sessions.

A complementary study of only the correct responses confirmed these findings. Mean SNR for A_2 and A_3 are nearly the same (Table III). But it would be misleading to conclude that these two systems present the same degree of intelligibility, for the distribution of erroneous responses indicates that text-to-speech systems are less intelligible than coded speech (Table IV).

Systems	Mean SNR	sd
A_1	2.24	.894
A_2	9.04	2.446
A_3	9.00	2.960
A_4	12.72	2.837

Table III - Mean SNR at the IT (dB) as a function of systems, for correct responses.

One can also see that a training effect appears only for the "bad" synthetic speech. Moreover, the extent of the improvement from the first to the second session varies depending on the sentence. But it is worth noting that some easy-to-parse sentences are less well understood than more difficult ones. These two results suggest that acoustic-phonetic cues play a role as well as syntactic or semantic information.

	Sessions		
Systems	1	2	Change
A_1	3	4	- 1
A_2	14	9.5	+ 4.5
A_3	26	23	+ 3
A_4	28	18.5	+ 9.5

Table IV - Percent erroneous responses, for each system and each session.

Analysis of perceptual confusions revealed systematic errors only for the text-to-speech systems. For example, initial /m/ and the nasal opposition, initial /v/ and the opposition /v-f/ led to numerous identification errors. On the contrary, no systematic error appeared for LPC and natural speech. From a morpho-syntactic and syntactic point of view, monosyllabic pronouns and prepositions were well identified, whereas mono or polysyllabic nouns in a subject noun phrase brought about errors. For all the positions, adverbs and adjectives were often omitted or modified. Generally speaking, errors located at the beginning of a sentence were usually not corrected, irrespective of syntactic structure. The only syntactic structure that was misunderstood was of the injonctive type (3 sentences). On the other hand, semantic plausibility played a role only when it was very low, irrespective of speech quality.
Though the verb is generally considered as the main component of a sentence /6/, the large number of misleading identifications of the first lexical items suggests that listeners processed information from left to right, according to a step-by-step decoding strategy. Sequential processing did not prevent backward error rectifying in some cases /4/. However backward corrections occurred just when first responses had exhibited a good degree of approximation to the signal. Fruitful corrections were always supported by a correct identification of sentence "scaffolding" provided by pronouns and prepositions. These results correspond to what can be called a comprehension strategy: locating syntactic marks allows the listener to restore missing phonemic or syllabic information. But a striking result of error analysis is that this kind of restoration appears only either for natural speech and coded speech or during the last trials for the other systems i.e. when the noise was very weak. The choice of a comprehension strategy is constrained by listening conditions as well as by the quality of signal. SNR analysis agrees with this assumption.
Intelligibility per se of the signal was evaluated for correct responses as a function of SNR at the identification threshold. In spite of some restrictions related to the range of syntactic structures /5/, cumulative frequencies for each SNR are a reliable indicator of intelligibility /10/. Identification data for each system are shown in Figure 1 as percentages of correct responses according to SNR carried out in each condition.

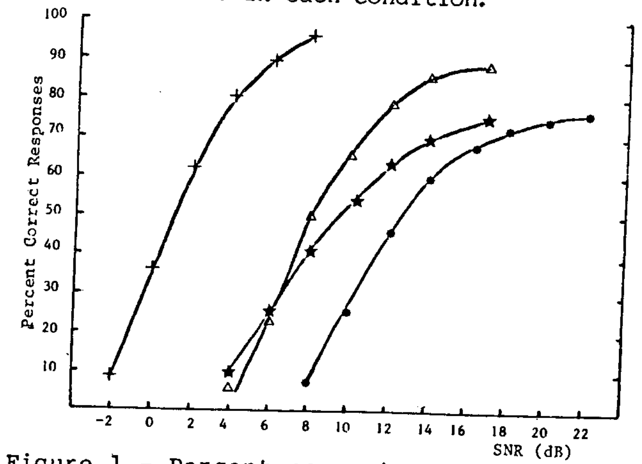

Figure 1 - Percent correct responses as a function of SNR, for each system. A_1: + ; A_2:△ A_3:★ ; A_4: ●.

- Intelligibility gain (IG) for natural speech reaches about 13% for a SNR of -2dB to +2dB, improves by 9%/dB between SNR of +2dB and +4dB, then stabilizes around 3%-4%/dB.
- As for coded speech, IG is 8%/dB between SNR of +4dB and +6dB, to 13%/dB between SNR of +6dB and +8dB, then decreases to 6%-8%/dB for a +8dB to +12dB SNR range, and to 3.6%/dB between SNR of +12dB-+14dB. A plateau is reached around a SNR of +18dB.
- A_3 high-quality text-to-speech system is characterized by a less steep gradient. IG varies from 8%/dB to 5%/dB between SNR of +4dB to +12dB, then increases slowly by 2.5%/dB till the highest SNR tested. Plausibly a plateau could appear around a SNR of 19dB or 20dB, and an intelligibility of 85% could be reached.
- Evolution of IG for the low-cost synthetic system A_4 is quite similar to that of A_3. The gain is rather strong at first (9%/dB to 7%/dB for a SNR range of +8dB to +14dB). It then decreases to 4%-2%/dB for a +14dB to +18dB SNR range. Around a SNR of +18dB, the slope flattens out.
The more striking feature of Figure 1 lies in the clear contrast between natural and synthetic speech intelligibility. Comparison between A_1 and A_2 shows clearly that intelligibility of LPC coded speech decreases as noise gets louder. Discrepancy between the two systems is maximum for a SNR of +4dB and reaches a 75% loss of intelligibility. This loss is then reduced to about 45%, but remains high even for a rather weak noise. Thus, in the best listening condition, coded speech intelligibility does not exceed 90%. This result agrees with the hypothesis bearing on the specificity of synthetic speech, compared with natural speech.
Secondly, asymmetry of A_2 and A_3 intelligibility curves gives some interesting information pertaining to listener's strategies. The

resistance to noise of A_3 is rather good for a loud noise. Intelligibility loss is indeed of 40% with regard to natural speech, but only of 10% compared with LPC speech. On the contrary, when SNR increases as noise becomes weaker, the gap between the systems A_2 and A_3 widens out. Assumption will be made that listeners adjust their identification strategy not only to the system, but also to the listening condition. When noise is very loud, they rely mainly on acoustic information. So very useful cues are given by the text-to-speech system. As a matter of fact, A_3 is characterized by clear segmentation cues, as for example prosodic cues i.e. F_0 movements and syllabic lengthening which are cues to word boundaries in French. As listening conditions get better, listeners can adopt another strategy, and give more attention to the sentence as a whole. This global comprehension strategy greatly improves the responses for a rather redundant speech as A_1 or even A_2, but it does not find a sufficient ground in impoverished synthetic speech to really succeed in A_4 and even A_3. That is perhaps why guessing or backward restoration very often fail when listeners are working with the two text-to-speech systems.

CONCLUSIONS

Impoverishment of speech by a pink noise varied mainly as a function of the system from which signal was generated. Relative weakness and lack of stability of sentence effect suggest that perceptual processing, in this experiment, has borne mainly on acoustic-phonetic cues, and secondly on prosodic segmentation cues. Listeners relied more on acoustic than on specifically linguistic information. Higher-order information was used, as demonstrated by the occurrence of backward lexical identification mechanisms; but its effect depends on the main effect of the quality of the system. Our results agree with the conclusion of Nusbaum and Pisoni /7/: "the differences in perception of natural and synthetic speech are largely the result of differences in the acoustic-phonetic structure of the signals" (p. 239). However, unlike them, we found that linguistic context becomes more important as the quality either of speech or of listening gets better, as is the case when one examines error restoration as well as identification thresholds. Furthermore, acoustic information is all the more processed as either speech quality or SNR are worse. In such bad listening conditions, subjects process the signal in a step-by-step fashion, more clearly so for synthetic speech than for natural speech.
Dissymmetry between responses is sufficient to rule out the hypothesis that synthetic speech is equivalent to natural speech degraded by noise. On the contrary, our results agree with the definition of synthetic speech as "impoverished speech" /7/, different in its nature from natural speech. They support the conclusion that the differences of intelligibility between natural and synthetic speech are related to the characteristics of speech signal. Different generating systems offer different patterns of cues to listeners. So listeners must construct and process several "perceptual spaces". The three synthetic speech systems generally present the same kind of confusion errors, more or less frequent depending on the quality of the system. Furthermore, two kinds of processing strategies can be hypothesized: a step-by-step decoding strategy and a global comprehension strategy. But further research is needed to better understand how perceptual spaces are built, what their consistency is, and how their processing can be improved.

REFERENCES

/1/ P. Combescure, 20 listes de dix phrases phonétiquement équilibrées, "Revue d'acoustique", 56, 34-38, 1981.
/2/ S.L. Greenspan, H.C. Nusbaum, D.B. Pisoni, "Perception of synthetic speech: Some effects of training and attentional limitations", Bloomington: Indiana University, Speech Research Laboratory, Progress Report, 387-413, 1985.
/3/ F. Grosjean, Spoken word recognition processes and the gating paradigm, "Perception and Psychophysics", 28, 267-283, 1980.
/4/ F. Grosjean, The recognition of words after their acoustic offset: Evidence and implications,"Perception and Psychophysics", 38, 299-310, 1985.
/5/ D.N. Kalikow, K.N. Stevens, L.L. Elliott, Development of a test of speech intelligibility in noise using sentence materials with controlled word predictibility, "J. of the Acoust. Soc. of America", 61, 1337-1351, 1977
/6/ G. Noizet, S. Bleuchot, R. Henry, Influence de la structure syntaxique de phrases entendues sur les stratégies de leur décodage perceptif, "Cahiers de Psychologie", 16, 149-180, 1973.
/7/ H.C. Nusbaum, D.B. Pisoni, Constraints on the perception of synthetic speech generated by rule, "Behavior Research Methods, Instruments & Computers", 17, 235-242, 1985.
/8/ D.B. Pisoni, Perception of speech: The human listener as a cognitive interface, "Speech Technology", 1, 10-23, 1982.
/9/ A. Salasoo, D.B. Pisoni, Interaction of knowledge sources in spoken word identification, "Journal of memory and language", 24, 210-231, 1985.
/10/ C. Sorin, Evaluation de la contribution de F_0 à l'intelligibilité, "Recherches/Acoustique", CNET, Vol. VII, 141-154, 1982/1983.